refArslan08

Self-timing schemes for SAE and WL driver controls signal generation rely on a BL-replica for giving global PTV information. However, a single replica bit-cell scheme can be fatal in today's scaled variation prone technologies. This is because the replica bit cell fabricated can be anywhere in the read current Gaussian curve, and doesn't correlate well with the ideal replica, which is meant to give the global mean read current.

This paper attempts to find a replica scheme that is more local variation tolerant. One way to get a global mean read current is to simply have numerous replica bit-cells on one or more replica bit-lines. This will average out the read current variation. However, this method doesn't scale well as the number of replica cells required rises sharply with increased variation. The approach used by the authors is to have a limited number of configurable replica bit cells, which can be programmed to act as replica cells driving the replica bit line, or simply acting as dummy loads. The replica bit cells programmed to act as actual replicas are thus a subset of the total number of configurable replica cells. This subset is chosen post fabrication. The method of selection involves first measuring the mean read current amongst all the possible subsets. Thus the subset closest to the mean is selected as the subset acting as actual replicas. Thus, this subset's total read current is proposed to be closely tracking the global mean read current (called the reference current) of an array bit-cell.

It is mentioned that the absence of an accurate reference current is a drawback of this approach. The potential benefits in terms of area, power, leakage and performance are all favorable.

refAmrutur01

Decoder design challenges include power and performance optimization. This is because decoder lies in the critical path of the read cycle. The large size of modern memory mean that decoders are large, and there is a need to find optimal number of stages, fan-in, fan-out, sizing strategies and circuit topologies for basic gates.

The approach taken by this work is that they use RC models of basic gates to solve analytically for the delay of a decoder. The stages selected are predecode-decode, with a parallel block selection decoder. In the analytical derivation the RC of interconnect is also included. The optimal bound on the delay of the decoder is found, this is used as a bench-mark for the decoder options later discussed.

For decoder basic gates, skewing is chosen a low power method of improving performance. However, this means that a separate reset path is needed in each skewed gate. This means extra area. Different resetting gates are discussed, and separate solutions are selected for different parts of the decoder. For the word-line drivers no skewing is done as number of word-line drivers is large and this means lot of area penalty. Finally, a predecoder with FI of 3, decoder stage2 and word line drivers with FI of 2 are decided upon.

refLai09

The background for this paper is the same as refArslan08, i.e. the inability of a single replica bit-cell to accurately track the global mean read current of the array.

This paper uses the BIST of a memory to measure the read delay of each bit-cell, and finds the worst case delay. This delay value is then converted into a digital code that is fed into a variable delay line. This delay line in put in the conventional replica bit line timing path, and adds to its delay. The code is calculated such that the total delay closely tracks the worst case bit-cell's read delay. This seems to be an efficient way to measure the reference global read current mean that refArslan08 talks about.

Since BIST is a regular feature of current day memories, the area overhead of this technique is nil. Further, by reducing the timing margins that are conventionally added to memory timing controllers, this method save s a lot of bit-line delay and power dissipation.

refAmrutur98

This is one of the first papers that proposes bit-line replica scheme for self-timed SRAM.

SRAM timing includes generation of the WL pulse and the SAE signal. The duration of the pulse, and the time at which SAE comes depends upon the amount of time it takes for sufficient amount of differential to develop on the BLs.

The simplest way to ensure correct functionality is to simulate and find the worst case local and global variation corner at the worst case temperature and voltage, and to factor enough margins in the WL pulse and SAE signal. However there are major drawbacks to this approach, namely performance and power penalties due to pessimistic swing of the bit line, additional power dissipation in sense amplifier, and can also lead to read upsets due to long WL pulse.

The approach to be less pessimistic is to somehow assess the global PTV cornet of the chip, and adaptively set the delay in the timing generation. Note that local process variation can still not be accounted for and worst case design has to be done with respect to local variation. The approach used by the author to sense the global PTV is to have a replica of the actual read circuitry and the bit-cell, and to see how much time that path takes to do a successful read. This time is then used as the delay for generating actual SAE and WL pulse signals.

The obvious assumption in this technique is that the replica bit-cell accurately tracks all the other bit-cells in the array, which in the current process nodes is far from the truth. refLai09 and refArslan08 propose to solve this problem.

refSamson

The challenges in decoder design in modern SRAM are speed and power, and correct timing with respect to the SAE signal. This paper highlights these issues.

It talks about the importance of clock and address line loading and how power consumption can be lowered by reducing these loads. The challenge is to lower these loads, while also being fast. The paper also talks about leakage reduction in decoder design.

The approach used by the authors is to reduce clock and address line loading by increasing the number of stages in the decoder. Speed of the decoder is still maintained by using dynamic gates instead of static gates. Further, excessive dynamic power dissipation, which is an issue in dynamic gates, is avoided by power gating unused decoder blocks. Power gating the unused blocks also helps save leakage power in decoder.